An Alternative Bilateral Refitting Model for Zooarchaeological Assemblages

Matthew O'Brien and Curtis B. Storlie

Abstract

Since the 1980's, the development of anatomical refitting methods opened the door to interpreting the single versus multiple occupations, separate households versus distinct activity areas, and unique food sharing of archaeological sites. In particular, bilateral refitting is a useful tool to link the social concepts and theory from cultural anthropology and apply them to the static remains of the archaeological record. Recently, critiques have raised concerns about the accuracy and precision of predictions that has limited the application of bilateral refitting. Bilateral asymmetry and large sample sizes have inhibited the success of univariate and bivariate refitting schemes. This paper presents a multivariate model that renews the potential of anatomical refitting. Through a battery of trials on simulated assemblages of pronghorn (Antilocapra americana) humeri, the results indicate significantly higher rates of successful matches and lower rates of Type I and Type II errors than existing methods.

1. Introduction

In the past thirty years faunal analysis has made significant strides in quantifying past social and subsistence activities. Through the use of experimental and ethnoarchaeological research, we have moved beyond pure descriptive to more behaviorally-based interpretations. Together with advances in quantitative and statistical software, anthropology has the opportunity to apply these tools toward old problems in archaeology. This paper addresses the topic of anatomical refitting of appendicular skeletal elements. In particular, we have developed a new flexible approach that incorporates multiple metric variables to allow for successful bilateral pairing within larger sample sizes.

The origins of zooarchaeological anatomical refitting dates back into the 1970s with the attempt to utilize the Lincoln Index in faunal analysis (see Lyman 2008:129). The methods employed on the Horner Site provided the first well-defined and replicable means of identifying potential refits within a zooarchaeological assemblage (Todd 1983). To address bilateral pairs, Todd used a series of measurements to capture the relative morphology of limb bones. Using a comparative collection of known pairs, a correlation test was run to isolate which of these measurements yielded the highest correlation coefficient. Using that single measurement on the archaeological sample, Todd was able to isolate potential matches. With the narrowed down list of possible matches, non-metric skeletal signatures were then used to identify whether the predicted match was accurate. For example, to find the correct match for a distal left tibia fragment, a single measurement from it would be compared to all right distal tibia specimen. Todd's interest was to understand the temporal relationship of multiple bone deposits. This univariate approach to identifying possible pairs was used to address food sharing (Waguespack 2002) and socioeconomic organization (Enloe 1991, 2003; Enloe and David 1992). While using different methods, Lyman (2006) and Adams and Konigsberg (2004) used bilateral refitting to approximate Lincoln Index MNI values.

Despite the utility of anatomical refitting, the existing methods typically lead to statistical errors that cloud the utility of refitting schemes. Lyman (2006, 2008) argues that existing refitting approaches are likely to produce Type I and Type II statistical errors. In reference to bilateral refitting, Type I Statistical errors would be the inability of a model to identify a particular specimen's bilateral pair. Type II errors refer to the identification of a match between two different individuals. An inverse relationship exists between Type I and Type II errors. A model with liberal parameters (i.e. thresholds defining a match) will result in more Type II errors, but fewer Type I errors. If non-metric signatures are ambiguous, then the Type II errors may be accepted as true pairs. Stricter parameters will produce fewer Type II errors, but more Type I errors. This means many skeletal elements would remain unpaired regardless if their actual match was within a given sample.

Of the two types of refitting errors, the Type II errors are more problematic within archaeological assemblages. A model that indicates a false positive can alter a researcher's interpretation of an archaeological site. In actualistic studies, there is often no way of identifying a false positive. These Type II errors can lead to misleading interpretations of the faunal assemblage. Type I errors are acceptable given the common condition of faunal remains. Typically, we only recover a small portion of the bones that were initially discarded. The impact of taphonomic processes leaves the analyst with an incomplete sample that will likely have appendicular elements with no bilateral pair within the sample. This suggests the presence of unmatched bones is an acceptable condition as long as this is tethered to a reduction in the number of false positives.

To target Type II errors, a model must be able to recognize the best match for both potential pairs. For example, Figure 1 represents a sample of left and right bones. If our goal is to find the best match for Bone A, then the easy choice is Bone B. If the process stops there, the model has likely identified a false positive. If the model also compares the best match for Bone B, we would find that Bone C is its best match. Obviously, Bone B and C are a better match, but how can a model select the appropriate match? The key is to run a model from the perspective of both sides (left and right) and choose the lower of the two probabilities. In this hypothetical case, the lowest probability between Bone B and C is still higher than the probability of Bone A and B from the perspective of Bone B. The model needs to operate along this path of logic to lower the chances of the Type II errors.

Sample size and asymmetry of bilateral pairs are the primary problems facing existing bilateral refitting schemes. Enloe (1991, 2003) and Lyman (2006, 2008) argue that sample size must remain low to prevent clustering of data that inhibits identification of bilateral pairs. Clustering refers to the overlap in measurements from separate individuals that pose as additional potential matches for a given specimen. Larger sample sizes will result in significant overlap that prevents clear indications of actual pairs. Within a given species, young and older individuals are more susceptible to mortality and therefore less common in the demographic profile of species. As a result, the demographic profile will tend to approximately have a normal distribution centered on young adults. In terms of refitting, a normal distribution leads to significant overlap in bone measurements as the number of individuals increase. This is alleviated in some species with strong sexual dimorphism, but this too will be problematic as the sample size continues to rise.

The second concern is asymmetry. It is well-established that our bones are not exactly the same bilaterally (Klingenberg et al. 2002; Leamy et al. 2001). While variation exists between bilateral pairs, the symmetry, or geometry, within a single element is consistent. For example, the variance between the length and width compared to length and depth of a single element will be similar. Slight variation exists in the morphology of our bones and depending on the severity of asymmetry, this factor could lead to both types of statistical errors. To address asymmetry, a new method must incorporate the existing variance between bilateral pairs.

In attempt to highlight the dilemma of asymmetry, Lyman (2006) analyzed white tailed and mule deer humeri and astragali. Unlike Todd's univariate approach, he used two measurements to identify bilateral pairs. Arguing that most studies of bilateral refitting assume bilateral symmetry, Lyman incorporated the two measurements and the Pythagorean theorem to quantify the amount of asymmetry.

$$\square=\surd(\square\square-\square\square2+ \square\square-\square\square2)$$

The test of astragali symmetry yielded variances that ranged from 0.347 mm to 0.561 mm. So is this too much variance? Even with the inclusion of two variables, his model could not deal with the data clustering that begins to occur with 17 white tailed deer astragali. The impact of the asymmetry issue alone is not clear, but when it is combined with increased sample sizes bilateral refitting is severely inhibited.

The range of anthropological applications suggests that we continue to pursue a methodology that can narrow down the potential bilateral refitting, but that we must address the pitfalls of previous approaches. Bilateral refitting is time consuming, but the successful identification provides a rare glimpse into past events that other streams of evidence cannot provide. This paper presents a new multivariate approach that increases the frequency of positive matches and minimizes the number of Type I and Type II errors relative to existing methods.

2. Material and Methods

To develop a new method of anatomical refitting, our model takes full advantage of the multivariate nature of measurements included to increase statistical power. To deal with the asymmetry problem, the approach also incorporates the variance existing between two known pairs into determining pairs within the test sample. We will first outline the basic structure of the statistical methodology and then introduce the comparative and test assemblages used to test the model.

2.1 Refit Model

In order to identify bilateral pairs in the presence of substantial data clustering, it is necessary to increase the number of measurement dimensions, or variables, per skeletal element. For this paper, we chose to use the measurements used by Todd (1983, 1987), Enloe (1991, 2003) and Waguespack (2002). Additional measurements are also possible, including those derived from the use of 3D laser scanning. If covariances and correlations are held constant, then the more variables that are used will result in more accurate results. Let $\boldsymbol{x}_i$, $i = 1,\ldots m,$ be the vector of measurements made on the i-th left skeletal element in the sample, $\boldsymbol{x}_i = [x_{1,i}, x_{1,i}, \ldots, x_{p,i}]'$, where p is the number of separate measurements made on each skeletal element. Similarly let $\boldsymbol{y}_j$, $j = 1,\ldots,n$ denote the vector of measurements made on the j-th right skeletal element in the sample. To calculate the probability of a refit, we make use of a multivariate normal model for the difference ***d*** between two corresponding ($i \leftrightarrow j$) right and left measurements, $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, respectively. Namely, for a corresponding pair ($i \leftrightarrow j$), we assume

$$\boldsymbol{d} = (\boldsymbol{x}_i - \boldsymbol{y}_j) \sim N(0,\Sigma),$$

where $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ represents the multivariate normal distribution with mean vector $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. While we acknowledge the presence of asymmetry within bilateral pairs, the model assumes that there is no difference in the size of bilateral pairs on average (i.e. $\boldsymbol{\mu} = 0$). This is to say that we would not *expect* a measurement on a left skeletal element to be greater than the same measurement on the corresponding right skeletal element and vise versa. Simply put, left Bones are not systematically bigger than right bones, and vise versa. Under this model, we can calculate the probability that $i \leftrightarrow k$ given that $i \leftrightarrow j$, for some $j = 1,\ldots,n$. That is, if we assume that there **is** a refit for the i-th left skeletal element in our sample, then we can calculate the probability that it corresponds to a particular (k-th) right skeletal element. Let $\boldsymbol{d}_{ij} = \boldsymbol{x}_i - \boldsymbol{y}_j$, and we can write this conditional refit probability λ_{ik} formally as

$$\begin{aligned} \lambda_{ik} &= \Pr(i \leftrightarrow k \mid i \leftrightarrow j, \text{ for some } j = 1,\ldots,n) \\ &= \Pr(\boldsymbol{d} = \boldsymbol{d}_{ik} \mid \boldsymbol{d} = \boldsymbol{d}_{ij}, \text{ for some } j = 1,\ldots,n) \qquad (1) \\ &= \frac{\phi(\boldsymbol{d}_{ik};0,\Sigma)}{\sum_{j=1}^{n} \phi(\boldsymbol{d}_{ij};0,\Sigma)}, \end{aligned}$$

where $\phi(\boldsymbol{d};\mu,\Sigma)$ is the multivariate normal density function (Johnson and Wichern 2003:143 [4-11]),

$$\phi(\boldsymbol{d};\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{d}-\mu)'\Sigma^{-1}(\boldsymbol{d}-\mu)\right\}. \qquad (2)$$

See the Appendix for a formal derivation of (1). The Refit Probability λ_{ik} provides a measure of the likelihood of *i* (a particular left sided bone) matching *k* (a particular right sided bone) according to the observed measurements. We can also calculate the conditional refit probability ρ_{jl} for a particular right bone *j* to a particular left bone *l*, given that there is a match for the *j*-the right bone in the sample. This is just the mirror image of the calculation given in (1), namely

$$\begin{aligned}\rho_{jl} &= \Pr(j \leftrightarrow l \,|\, j \leftrightarrow i, \text{ for some } i = 1,\ldots,m)\\ &= \Pr(\boldsymbol{d} = \boldsymbol{d}_{lj} \,|\, \boldsymbol{d} = \boldsymbol{d}_{ij}, \text{ for some } i = 1,\ldots,m)\\ &= \frac{\phi(\boldsymbol{d}_{lj};0,\Sigma)}{\sum_{i=1}^{n}\phi(\boldsymbol{d}_{ij};0,\Sigma)},\end{aligned}$$

In order to calculate λ_{jk} and ρ_{jl} on a test sample of unknown individuals, we need to specify the unknown covariance matrix $\boldsymbol{\Sigma}$. This can be done by substituting in the maximum likelihood estimates based on a sample of $\boldsymbol{d}_{ij}$ from which the $i \leftrightarrow j$ relationships are known.

The exponent of the numerator in equation (2) is the Mahalanobis Distance between $\boldsymbol{x}_i$ and $\boldsymbol{y}_k$. This effectively quantifies the amount of asymmetry between $\boldsymbol{x}_i$ and $\boldsymbol{y}_k$. In a perfectly symmetrical sample, for $i \leftrightarrow k$ the difference $\boldsymbol{d}_{ik} = \boldsymbol{x}_i - \boldsymbol{y}_k$ would be equal to zero, which means that the Mahalanobis Distance would equal zero. For the purposes of bilateral refitting, $\phi(\boldsymbol{d};0,\Sigma)$ represents the coefficient of asymmetry within a given skeletal element within a single species. For any given test sample, the $\phi(\boldsymbol{d};0,\Sigma)$ will be unique, depending on the vector of measurements taken and the species, etc.. Unlike previous refit schemes, this approach does not "pick" a value of asymmetry, which then establishes the threshold to accept or reject a statistical refit.

Based on the multivariate density function and our refit probability function, we built a working model using R version 2.9, which is open-source statistical software (http://cran.r-project.org/). Two probability matrices are constructed: one matching from the perspective of the sample of left bones, $\{\lambda_{jk}, i = 1,\ldots,m, k = 1,\ldots,m\}$, and the second from the perspective of right bones, and $\{\rho_{jl}, j = 1,\ldots,n, l = 1,\ldots,n\}$. The results are then tabulated into two matrices: a minimum probability matrix $P_{\min}$, whose i-th row, j-th column is $P_{\min,ij} = \min(\lambda_{ij} , \rho_{ji})$ and a maximum probability matrix, $P_{\max}$ whose i-th row, j-th column is $P_{\max,ij} = \max(\lambda_{ij} , \rho_{ji})$. The minimum probability matrix ($P_{\min}$) will yield the more cautious results by reflecting the lower of two probabilities to minimize Type II errors. The maximum probability matrix ($P_{\max}$) reports the higher of the results to maximize the number of positive matches and minimize the number of Type I errors. The drawback of the second matrix is that it will likely cause more false positive results. In practice, these measures should be used as a lower and upper bound, respectively, on the likelihood of a match between skeletal elements i and j. Any pairs (i,j) that have a high enough value (above some threshold T) for $P_{\min,ij}$ and/or $P_{\max,ij}$ could be chosen as candidates for further analysis. The actual value of T used in a particular analysis will be dependant on the number of matches that can feasibly be followed up with further non-parametric analyses. The value of T could also be set to achieve a desired type II error by using cross validation on the test sample.

2.2 Data Selection

The refit model requires two independent samples to operate. The first sample is a comparative sample of known individuals that can be used to establish the covariance matrix. The second sample of individuals is a test sample on which to evaluate potential matches. In this presentation, the relationships in the test sample are also known so that we can evaluate the success of the proposed approach. In practice the relationships in the test sample would not be known of course, and hence the need for the proposed approach. The species used is pronghorn (*Antilocapra americana*). This study uses eight post-cranial remains from the University of Wyoming's Zooarchaeological Lab and nine individuals housed at the University of New Mexico's Museum of Southwest Biology. As previously mentioned, this analysis mirrored the metrics from Todd (1983). Each measurement was taken three separate times using digital calipers accurate to +/- 0.3 mm. The averages of those three measurements were taken as the estimated length for the purposes of the model. This paper explores the functionality of the refit model by using only the distal portion of pronghorn humerus. In total, Todd established six separate measurements (i.e., $p = 6$) for this portion of the skeleton (**Table 1**) (1983, 1987). Each distal humerus was measured 18 times in total for a total of 306 measurements. For complete bones, analysts following Todd's methods can collect up to 15 different measurements from each specimen, but this is time consuming and often not practical given the problems of weathering and post-depositional processes that break down zooarchaeological assemblages.

The goal of this research is to identify the effectiveness of this approach with larger samples. In order to do this, a simulated assemblage of humeri was randomly generated using a multivariate normal distribution with parameters obtained from the MLE estimates of the comparative sample. Specifically, we assume that the vector $\boldsymbol{z} = [\boldsymbol{x}_i, \boldsymbol{y}_j]$, of 12 measurements from a corresponding pair of bones (six measurements on left bone and right bone, respectively) follows a multivariate normal distribution, i.e., $z \sim N(\mu_z, \Sigma_z)$, where

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \; \Sigma_z = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}' & \Sigma_y \end{bmatrix}. \tag{3}$$

It is assumed that the distribution of $\boldsymbol{x}_i$ is the same as that of $\boldsymbol{y}_j$, i.e., there is no systematic difference between left bones and right bones as mentioned earlier. Therefore, the model in Equation (3) is restricted such that $\mu_x = \mu_y$ and $\Sigma_x = \Sigma_y$. The resulting MLEs under this model using the comparative sample are provided in Table 2. Notice the positive values in the Σ_{xy} matrix. The primarily positive correlations, especially on the diagonal of the Σ_{xy} matrix are what yield the discernment power of the proposed method.

Given the model in Equation (3), we can generate a test sample to evaluate the proposed method with the `mvrnorm` function in R. Specifically, using the μ_z and Σ_z in Table 2, the command

```
test.sample <- mvrnorm(n, mu.z, Sigma.z)
```

will generate a n × 12 matrix, each row of which is a sample of corresponding left and right bones (the first six measurements of each row correspond to a the left bone, and the last six to the right bone, of the same individual).

A sample of the generated measurements is provided in Table 3. A total of fifteen simulated pronghorn humerus test samples were used for this analysis that ranged from $n = 10$ to 50 individuals. While it is possible to examine the impact of larger sample sizes, the majority of archaeological assemblages fall within this range of individuals. The organization of these samples is described in more detail in the next section.

3. Results

The refitting model was run under a series of different tests. An arbitrary threshold of (P ≥ 0.85) was selected to determine a match. A probability equal to or greater than the threshold was determined a match. In circumstances that more than one match exceeded the threshold, the highest probability was chosen to be the correct match. The summary tables for each diagnostic tests provide the number of correct matches, Type I errors (T I), Type II errors (T II), and their respective percentages. In practice, all specimens that exceed the threshold should be inspected for non-parametric attributes to identify the best fit. The first trials ran the model against various random samples of increasing size. These samples included each of the left and right matches for each individual – equal numbers of left and right humeri. This first test also ran Lyman's (2006) method against the same samples for a comparison of results. The second test altered the number of measurements per humerus to identify how fewer variables impact the refitting model. Finally, the test examined how the inclusion of uneven numbers of lefts and rights would impact the accuracy of the model.

The first test of the model examined the impact of increasing sample sizes on the reliability of matching an individual specimen's bilateral elements. All tests were run with the maximum six variables available for the distal humerus. Random sample sizes of 10, 20, 30, 40, and 50 individuals were generated for three separate trials to approximate the model's effectiveness. HM11 and HM14 were the two measurements used for Lyman's (2006) bivariate model based on bilateral correlation values (**Table 4**). The conservative matching statistic (c) was 0.36 and the liberal value was 0.52. Both values and their associated results are presented.

The results of the variable sample size test show P_{min} identifies a match on average 46 percent of the time (**Table 5**). More importantly, false positives only occur 4 percent of the time on average. Using P_{max}, the average rate of identifying a pair increases to 79 percent with over a 9 percent chance of a Type II error. To see the impact of sample size more clearly, a pair of plots shows the decline of correct matches as the sample size increases (**Fig. 2**). P_{min} has an almost linear inverse relationship with sample size, which adheres to Lyman's (2006, 2008) and Enloe's (1991, 2003) predictions. Whereas the success rate of identifying the correct match occurs over 80 percent of the time with samples of 10 individuals, this percentage drops to an average of 34 percent with samples of 50 individuals. The impact of sample size is less severe in the P_{max}, which perfectly matched all individuals with 10 individuals, and remained successful nearly 80 percent of the time with 50 individuals. Of possibly greater importance, P_{min} is more successful at minimizing Type II errors than P_{max}. Regardless of sample size, the average of Type II errors in P_{min} remains below 5 percent. P_{max} is less accurate at minimizing false positives, but the average rate is less than 10 percent. This data suggests that the model adequately identifies pairs with minimal chances of misidentification when both matrices are used in conjunction.

The results from Lyman's model fits with his assumption of increasing sample sizes. As sample size increases, the number of correct matches declines rapidly and the frequency of Type

II errors increases significantly (**Table 6**). **Figure 3** shows the inverse relationship of correct matches in relation to sample size and the positive relationship between false positives and the increase in individuals within a sample. Using the conservative threshold, correct matches occur at slightly greater than 20 percent, while Type II errors occur 30 percent of the time. The liberal threshold increases the likelihood of a correct match, but results in a marked increase in the number of false positives.

The second series of tests examined the impact of reducing the number of variables used to distinguish pairs. We used the three trials of 20 individuals from the previous test. The bilateral correlations derived from the comparative sample were used to decide which measurements were removed first. From five down two variables, each of the following measurements was removed in this order: HM8, HM7, HM15, and HM6.

On average, the results from the second test of the model fall in line with logical expectations (**Table 7**). As the number of variables decrease, the effectiveness of identifying matches decreases. With normally distributed data, the reduction in variables results in less effective identification of correct matches. The P_{min} matrix prevents an associated increase in false positives as the number of variables decreases. P_{max} is less successful at minimizing Type II errors, but identifies correct matches with greater frequency. Despite the rise in the frequency of Type II errors, they still fall well below the results from Lyman's model for a sample of 20 individuals.

The final test for the model was to generate hypothesized archaeological assemblages with uneven preservation or presence of an individual specimen's bilateral pair. The goal is to test the accuracy of the model when there is no match for a portion of the specimen. Since the entire test sample is random, there was no need to randomize the selection process for this test. Humeri were arbitrarily removed from a randomly generated complete pairs sample to create the wanted quantity of left and right bones.

Table 8 provides the difference in the number of left to right humeri and the results of the test. Overall, the $P_{\min}$ predicts correct matches in slightly less than 50 percent of the individuals and limits the number of Type II errors occur approximately 10 percent of the time. In samples of 40 individuals or less, Type II errors occur on average of 7 percent. $P_{\max}$ maintains a success rate exceeding 77 percent, but the likelihood of Type II errors increases to 12.5 percent. The uneven distribution of lefts and rights had a mixed impact on the model's ability to predict pairs. In some cases the matching success rates stay about the same as in the even distribution of sided element trials, but most saw a decline in the number of successful matches. Most important, the $P_{\min}$ still produced percentages of Type II errors that were less than 10 percent.

Discussion

The results of this analysis indicate that it is possible to identify individuals within large samples if the analysis uses additional metric dimensions. The importance of using multiple measurements decreases the frequency of Type II errors and increases the likelihood of correct pairings. Despite its success in identifying bilateral pairs, results from empirical studies still need to be physically verified to confirm matches. For additional confidence, analysts should also consult other zooarchaeologists to confirm identified matches.

A primary issue with archaeological assemblages is the degree of weathering that limits the number of accurate measurements that can be captured per specimen. The diagnostic tests indicate that fewer variables decrease the model's ability to distinguish pairs, but the probability of Type II errors remains low. This is an important aspect of the model. When identifying matches with heavily weathered assemblages, the chances of a match will be contingent on the number of measurements that can be reliably recorded. When an analyst is limited to only two variables, the model will not be able to overcome data clusters. Instead of forcing a match, the model will opt to make a Type I error instead. This provides additional assurances to the reliability of a potential match, or matches, predicted by this approach.

In comparison with Lyman's bilateral refitting model, our approach provides improved rates of pair identification and reduced frequencies of Type II errors regardless of the number of individuals or variables. According to the diagnostic tests of equal numbers of left and right bones, the associated regression equation for P_{min} is:

Probability of a Correct Match = 0.916 – 0.1222 (# of Individuals).

Overall, the conditional probability of the model generating a Type II error for a given specimen is 2 percent using P_{min} and 8 percent in P_{max}. This indicates that an analyst can be confident in matches produced from the $P_{\text{min},}$ 98 percent of the time. In the case of P_{max}, the analyst can still be certain that a match generated by the model is an actual pair 92 percent of the time when its actual match is present in the sample. In the case of uneven representation of left and right bones, the regression equation for P_{min} ($r^2 = 98.9$; p value = 0.006) and P_{max} ($r^2 = 69.6$; p value = 0.166) is:

[P_{min}] Prob. of a Correct Match = 0.808 – 0.114 (# of individuals) and

[P_{max}] Prob. of a Correct Match = 0.934 – 0.005 (# of individuals).

For a given specimen, the analyst can be confident that a predicted match using P_{min} is from a single individual 94 percent of the time. This confidence drops when using P_{max} to 89 percent. Using Lyman's conservative approach, the probability of a identifying a match is 0.52, but any identified match has a 16 percent chance of being a false positive. The results are worse with the liberal approach, which leads to a 0.75 probability of a match, but there is a 31 percent chance of a false positive.

Conclusions

The time investment involved with bilateral refitting is prohibitive with most faunal assemblages. In large assemblages, the total number of measurements needed for this approach can be excessive. Conservatively, this model should be reserved for archaeological sites that have well preserved faunal remains and well-preserved spatial context. In these circumstances, the spatial distribution of individual animal remains across a site can provide a unique view into site formation, spatially segregated activities, and/or the social interaction between different households. Analysts examining spatially segregated faunal assemblages within a site, or closely linked sites, could use bilateral refitting to identify whether individual animals were dismembered in a single location or processed in a series of stages located in different portions of a campsite. Researchers can also use bilateral refitting to identify single versus multiple occupations at a site. In a similar vein, the successful application of bilateral refitting can also help address identifying mass kill versus accretionary kill events. Finally, the application of bilateral refitting can be used to identify directional trends of past sharing behavior, which can be used in conjunction with Behavioral Ecological models (i.e. Waguespack 2002) to better understand transitions in social interaction over time.

The strength of this model lies in its flexibility to incorporate as many variables that can be reliably collected by the faunal analyst. It can be used in traditional analyses using digital calipers as well as more recently accepted use of 3D scanned images. This allows the model to be applied to existing faunal data as well as newly compiled data. Unlike many other advances in methodology or analysis, this model requires no investment in cost-prohibitive software or hardware.

Previous attempts to identify bilateral pairs raised concerns that large sample sizes lead to statistical Type I and Type II errors (Enloe 2003; Lyman 2006, 2008). Overlapping is common in skeletal measurements because of bone sizes are generally normally distributed. Our model presents an alternative approach to identifying bilateral pairs within the appendicular skeleton that increases the frequency of correct matches and limits the number of false positives. As the sample size increases, the conservative model (P_{min}) results in lower percentage of correct matches and Type II errors, while increasing the frequency of Type I errors. Inversely, the P_{max} results in lower percentage of Type I errors, but increases the likelihood of a correct match and Type II errors. Using the combination of P_{min} and P_{max} can provide the respective lower and upper boundaries of likely matches for a given bone, which can minimize the number of specimen that are physically inspected.

Appendix

Here we formally demonstrate the relation in equation (1). We need to show that for a random vector ***d*** with multivariate normal distribution $N(\mathbf{0},\boldsymbol{\Sigma})$,

$$\Pr(\boldsymbol{d} = \boldsymbol{d}_{ik} \mid \boldsymbol{d} = \boldsymbol{d}_{ij}, \text{ for some } j = 1,\ldots,n) = \frac{\phi(\boldsymbol{d}_{ik};0,\Sigma)}{\sum_{j=1}^{n}\phi(\boldsymbol{d}_{ij};0,\Sigma)}.$$

Now, in general for some random vector $\boldsymbol{X} = [X_1, \ldots, X_p]'$ with a continuous multivariate cumulative distribution function (CDF) $\Phi(\boldsymbol{x})$, like the multivariate normal distribution, the mixed derivative of $\Phi(\boldsymbol{x})$ gives the density function $\phi(\boldsymbol{x})$, i.e.,

$$\phi(\boldsymbol{x}) = \frac{\partial^p}{\partial x_1 \cdots \partial x_p}\Phi(\boldsymbol{x}).$$

In the univariate case this leads to

$$\phi(x) = \frac{\partial}{\partial x}\Phi(x) = \lim_{\varepsilon \to 0}\frac{\Phi(x) - \Phi(x+\varepsilon)}{\varepsilon} = \lim_{\varepsilon \to 0}\frac{\Pr\left(x \le X \le x + \varepsilon\right)}{\varepsilon}.$$

In the multivariate case, this becomes

$$\phi(\boldsymbol{x}) = \frac{\partial^p}{\partial x_1 \cdots \partial x_p}\Phi(\boldsymbol{x}) = \lim_{\varepsilon \to 0}\frac{\Pr\left(\bigcup_{k=1}^{p}\left\{x_k \le X_k < x_k + \varepsilon\right\}\right)}{\varepsilon^p}.$$

Finally,

$$\begin{aligned}
\Pr(\boldsymbol{d} = \boldsymbol{d}_{ik} \,|\, \boldsymbol{d} = \boldsymbol{d}_{ij}, \text{ for some } j = 1,\ldots,n) &= \Pr\left(\boldsymbol{d} = \boldsymbol{d}_{ik} \,\middle|\, \bigcup_{j=1}^{n} \left\{ \boldsymbol{d} = \boldsymbol{d}_{ij} \right\} \right) \\
&= \lim_{\varepsilon \to 0} \Pr\left(\bigcup_{l=1}^{p} \left\{ d_{l,ik} + \varepsilon \le d_l \le d_{l,ik} \right\} \,\middle|\, \bigcup_{j=1}^{n} \left\{ \bigcup_{l=1}^{p} \left\{ d_{l,ij} + \varepsilon \le d_l \le d_{l,ij} \right\} \right\} \right) \\
&= \lim_{\varepsilon \to 0} \frac{ \Pr\left(\bigcup_{l=1}^{p} \left\{ d_{l,ik} + \varepsilon \le d_l \le d_{l,ik} \right\} \right) \Big/ \varepsilon^{p} }{ \sum_{j=1}^{n} \Pr\left(\bigcup_{l=1}^{p} \left\{ d_{l,ij} + \varepsilon \le d_l \le d_{l,ij} \right\} \right) \Big/ \varepsilon^{p} } \\
&= \frac{ \phi(\boldsymbol{d}_{ik}; \boldsymbol{0}, \Sigma) }{ \sum_{j=1}^{n} \phi(\boldsymbol{d}_{ij}; \boldsymbol{0}, \Sigma) },
\end{aligned}$$

which is the relation used in equation (1).

References Cited

Adams, B.J. and L. W. Konigsberg
2004 Estimation of the most likely number of individuals from commingled human skeletal remains. *American Journal of Physical Anthropology* 125:138-151.

Enloe, J. G.
1991 *Subsistence Organization in the Upper Paleolithic: Carcass Refitting and Food Sharing at Pincevent*, Doctoral dissertation, University of New Mexico.
2003 Food Sharing Past and Present: Archaeological Evidence for Economic and Social Interaction, *Before Farming* 1: 1-23.

Enloe, J. G. and F. David
1992 Food Sharing in the Paleolithic: Carcass Refitting at Pincevent, in: J. L. Hofman and J. G. Enloe (Eds.), *Piecing Together the Past: Applications of Refitting Studies in Archaeology*, British Archaeological Reports International Series 578, Oxford, pp. 296-299.

Johnson, R. A. and D. W. Wichern
2003 *Applied Multivariate Statistical Analysis, 6th* edition. New York, Prentice Hall.

Klingenberg, C.P., M. Barluenga, A. Meyer. Shape analysis of symmetrical structures: quantifying variation among individuals and asymmetry, *Evolution* 56: 1909-1920.

Leamy, L.J., S. Meagher, S. Taylor, L. Carroll, and W. K. Potts. Size and fluctuating asymmetry of morphometric characters in mice: their associations with inbreeding and the t-haplotype, *Evolution* 55:2333-2341.

Lyman, R. L.
2006 Identifying bilateral pairs of deer (*Odocoileus* sp.) bones: How symmetrical is symmetrical enough?. *Journal of Archaeological Science* 33:1256-1265.
2008 *Quantitative Paleozoology*. Cambridge. Cambridge University Press.

Todd, L. C.
1983 *The Horner Site: Taphonomy of an early Holocene Bison Bonebed*, Doctoral dissertation, University of New Mexico.
1987 Bison Bone Measurements, In *The Horner Site: The Type Site of the Cody Cultural Complex*, edited by G. Frison and L.C. Todd, Academic Press, New York: 371-403.

Waguespack, N. M.
2002 Caribou Sharing and Storage: Refitting the Palangana Site, *Journal of Anthropological Archaeology* 21: 396-417.

Tables and Figures

1. Tables

Table 1: Definition of distal humerus measurements defined by Todd (1983)

Measurements	Definition
HM6	Greatest Breadth of the Distal End
HM7	Breadth of Distal Articular Surface
HM8	Least Breadth of Olecrannon Fossa
HM11	Greatest Depth of Medial Distal End
HM14	Least Depth of Medial Distal End
HM15	Depth of Olecrannon Fossa

Table 2: The mean and variance of the metric variables

Simulated Sample Parameters	HM6	HM7	HM8	HM11	HM14	HM15
Measurement Mean	35.71	35.70	14.11	30.51	24.11	8.25
Measurement St. Dev.	2.22	2.23	1.21	1.41	1.54	0.78
Quotient Mean Based on HM6 (HM_X/HM6)	1.00	1.00	0.40	0.86	0.68	0.23
Quotient St. Dev. Based on HM6 (HM_X/HM6)	0.00	0.02	0.02	0.02	0.02	0.01
Assymetry Mean	0.32	0.47	0.26	0.22	0.25	0.17
Assymetry St. Dev.	0.26	0.41	0.22	0.16	0.14	0.14

Table 3: An example of simulated pronghorn humerus sample for trial one for sample size of 10 individuals

Individual	Side	HM6	HM7	HM8	HM11	HM14	HM15
1	Left	37.56	38.37	14.71	31.01	25.64	8.40
2	Left	35.25	35.41	13.60	30.21	22.58	9.12
3	Left	36.62	35.10	13.26	30.81	24.52	8.70
4	Left	34.28	34.50	14.08	29.83	22.11	8.02
5	Left	36.28	35.36	14.59	32.17	24.77	7.79
6	Left	39.91	40.48	17.05	34.61	28.20	10.17
7	Left	39.25	39.92	15.78	33.53	26.57	8.59
8	Left	33.30	33.54	12.36	27.13	22.31	7.99
9	Left	36.43	35.68	15.75	30.75	24.74	8.35
10	Left	34.60	35.38	14.00	28.86	23.04	8.45
1	Right	37.33	39.08	15.17	31.62	26.04	8.22
2	Right	34.75	35.59	13.45	29.88	22.16	8.66
3	Right	36.48	34.23	13.04	30.48	24.14	8.60
4	Right	34.13	33.83	13.94	29.80	21.92	7.47
5	Right	36.53	34.94	14.82	31.92	24.94	7.61
6	Right	39.20	40.00	16.61	34.06	27.99	9.72
7	Right	38.51	39.29	15.09	32.76	26.24	8.32
8	Right	32.68	33.03	12.01	27.17	21.87	8.08
9	Right	37.03	35.56	15.96	31.11	25.14	8.38
10	Right	35.03	35.74	14.18	29.43	23.03	8.65

Table 4: Correlation coefficients for distal humerus

Measurement	r
HM6-HM6	0.984
HM7-HM7	0.965
HM8-HM8	0.964
HM11-HM11	0.985
HM14-HM14	0.987
HM15-HM15	0.967

Table 5: Results from various sample sizes with the accepted matching probability threshold set at 0.85

		$P_{\min}$						$P_{\max}$					
Sample Size	Trials	Correct	T I	T II	%Correct	%T I	% T II	Correct	T I	T II	%Correct	%T I	% T II
10	1	10	0	0	100.0%	0.0%	0.0%	10	0	0	100.0%	0.0%	0.0%
	2	8	2	0	80.0%	20.0%	0.0%	10	0	0	100.0%	0.0%	0.0%
	3	8	2	0	80.0%	20.0%	0.0%	10	0	0	100.0%	0.0%	0.0%
20	1	13	6	1	65.0%	30.0%	5.0%	17	2	1	85.0%	10.0%	5.0%
	2	13	7	0	65.0%	35.0%	0.0%	19	1	0	95.0%	5.0%	0.0%
	3	10	10	0	50.0%	50.0%	0.0%	18	1	1	90.0%	5.0%	5.0%
30	1	16	11	3	53.3%	36.7%	10.0%	23	0	7	76.7%	0.0%	23.3%
	2	10	19	1	33.3%	63.3%	3.3%	23	5	2	76.7%	16.7%	6.7%
	3	20	8	2	66.7%	26.7%	6.7%	26	1	3	86.7%	3.3%	10.0%
40	1	12	26	2	30.0%	65.0%	5.0%	30	6	4	75.0%	15.0%	10.0%
	2	21	18	1	52.5%	45.0%	2.5%	31	7	2	77.5%	17.5%	5.0%
	3	19	20	1	47.5%	50.0%	2.5%	32	6	2	80.0%	15.0%	5.0%
50	1	13	32	4	26.0%	64.0%	8.0%	34	11	5	68.0%	22.0%	10.0%
	2	16	32	2	32.0%	64.0%	4.0%	33	8	9	66.0%	16.0%	18.0%
	3	22	27	1	44.0%	54.0%	2.0%	40	5	5	80.0%	10.0%	10.0%
450		211	220	18	46.9%	48.9%	4.0%	356	53	41	79.1%	11.8%	9.1%

Correct: The highest refit probability occurred between a single individual's left and right humeri

T I: Type I Error - The refit model failed to identify a match within the given sample

T II: Type II Error - The highest refit probability occurred between different individuals

Table 6: Refitting results using Lyman (2006) model against the first trial of each random sample

	Conservative (C) of 0.36						Liberal (C) of 0.52					
Sample Size	Correct	T I	T II	%Correct	%T I	%T II	Correct	T I	T II	%Correct	%T I	%T II
10	1	8	1	10.0%	80.0%	10.0%	3	5	2	30.0%	50.0%	20.0%
10	2	6	2	20.0%	60.0%	20.0%	3	5	2	30.0%	50.0%	20.0%
10	4	5	1	40.0%	50.0%	10.0%	6	3	1	60.0%	30.0%	10.0%
20	3	14	3	15.0%	70.0%	15.0%	7	9	4	35.0%	45.0%	20.0%
20	7	11	2	35.0%	55.0%	10.0%	13	4	3	65.0%	20.0%	15.0%
20	5	9	6	25.0%	45.0%	30.0%	8	5	7	40.0%	25.0%	35.0%
30	10	11	9	33.3%	36.7%	30.0%	12	5	13	40.0%	16.7%	43.3%
30	5	18	7	16.7%	60.0%	23.3%	5	8	17	16.7%	26.7%	56.7%
30	5	19	6	16.7%	63.3%	20.0%	7	14	9	23.3%	46.7%	30.0%
40	9	21	10	22.5%	52.5%	25.0%	15	7	8	37.5%	17.5%	20.0%
40	11	15	14	27.5%	37.5%	35.0%	17	5	18	42.5%	12.5%	45.0%
40	7	26	7	17.5%	65.0%	17.5%	9	13	18	22.5%	32.5%	45.0%
50	10	16	24	20.0%	32.0%	48.0%	15	3	32	30.0%	6.0%	64.0%
50	12	18	20	24.0%	36.0%	40.0%	18	8	24	36.0%	16.0%	48.0%
50	10	16	24	20.0%	32.0%	48.0%	18	6	26	36.0%	12.0%	52.0%
Total	101	213	136	22.4%	47.3%	30.2%	156	100	184	34.7%	22.2%	40.9%

Correct: The highest refit probability occurred between a single individual's left and right humeri

T I: Type I Error - The refit model failed to identify a match within the given sample

T II: Type II Error - The highest refit probability occurred between different individuals

Table 7: Impact of the number of variables using the three trials of 20 individuals from Table 5

		$P_{\min}$						$P_{\max}$					
Trial	Variables	Correct	T I	T II	%Correct	%T I	%T II	Correct	T I	T II	%Correct	%T I	%T II
1	6	13	6	1	65.0%	30.0%	5.0%	17	2	1	85.0%	10.0%	5.0%
	5	10	8	2	50.0%	40.0%	10.0%	15	5	0	75.0%	25.0%	0.0%
	4	5	13	2	25.0%	65.0%	10.0%	13	2	5	65.0%	10.0%	25.0%
	3	3	17	0	15.0%	85.0%	0.0%	9	8	3	45.0%	40.0%	15.0%
	2	0	20	0	0.0%	100.0%	0.0%	5	14	1	25.0%	70.0%	5.0%
2	6	13	7	0	65.0%	35.0%	0.0%	19	1	0	95.0%	5.0%	0.0%
	5	10	9	1	50.0%	45.0%	5.0%	17	0	3	85.0%	0.0%	15.0%
	4	8	11	1	40.0%	55.0%	5.0%	12	1	7	60.0%	5.0%	35.0%
	3	4	16	0	20.0%	80.0%	0.0%	12	6	2	60.0%	30.0%	10.0%
	2	4	16	0	20.0%	80.0%	0.0%	7	13	0	35.0%	65.0%	0.0%
3	6	10	10	0	50.0%	50.0%	0.0%	18	1	1	90.0%	5.0%	5.0%
	5	10	7	0	50.0%	35.0%	0.0%	17	1	2	85.0%	5.0%	10.0%
	4	7	13	0	35.0%	65.0%	0.0%	10	5	5	50.0%	25.0%	25.0%
	3	2	16	2	10.0%	80.0%	10.0%	9	5	6	45.0%	25.0%	30.0%
	2	0	20	0	0.0%	100.0%	0.0%	3	16	1	15.0%	80.0%	5.0%

Correct: The highest refit probability occurred between a single individual's left and right humeri

T I: Type I Error - The refit model failed to identify a match within the given sample

T II: Type II Error - The highest refit probability occurred between different individuals

Table 8: Uneven representation of sided elements

						$P_{\min}$						$P_{\max}$		
Left	Right	Trial	C	T I	T II	%C	%T I	%T II	C	T I	T II	%C	%T I	%T II
20	10	1	7	2	1	70.0%	20.0%	10.0%	8	0	2	80.0%	0.0%	20.0%
		2	7	3	0	70.0%	30.0%	0.0%	10	0	0	100.0%	0.0%	0.0%
		3	7	3	0	70.0%	30.0%	0.0%	8	1	1	80.0%	10.0%	10.0%
20	30	1	13	5	2	65.0%	25.0%	10.0%	16	0	4	80.0%	0.0%	20.0%
		2	9	9	2	45.0%	45.0%	10.0%	18	1	1	90.0%	5.0%	5.0%
		3	13	5	2	65.0%	25.0%	10.0%	18	0	2	90.0%	0.0%	10.0%
30	40	1	9	20	1	30.0%	66.7%	3.3%	22	3	5	73.3%	10.0%	16.7%
		2	15	12	3	50.0%	40.0%	10.0%	25	5	0	83.3%	16.7%	0.0%
		3	16	11	3	53.3%	36.7%	10.0%	16	11	3	53.3%	36.7%	10.0%
20	50	1	9	7	4	45.0%	35.0%	20.0%	13	2	5	65.0%	10.0%	25.0%
		2	8	8	4	40.0%	40.0%	20.0%	16	1	3	80.0%	5.0%	15.0%
		3	5	14	1	25.0%	70.0%	5.0%	15	1	4	75.0%	5.0%	20.0%
240			118	99	23	49.2%	41.3%	9.6%	185	25	30	77.1%	10.4%	12.5%

C: Correct - The highest refit probability occurred between a single individual's left and right humeri

T I: Type I Error - The refit model failed to identify a match within the given sample

T II: Type II Error - The highest refit probability occurred between different individuals

Figures

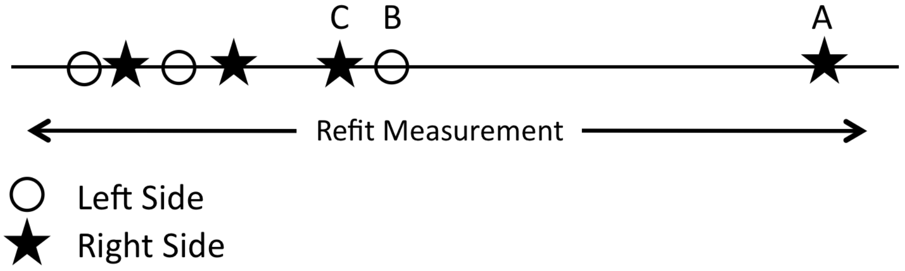

Figure 1: Importance of comparing left to right and right to left results

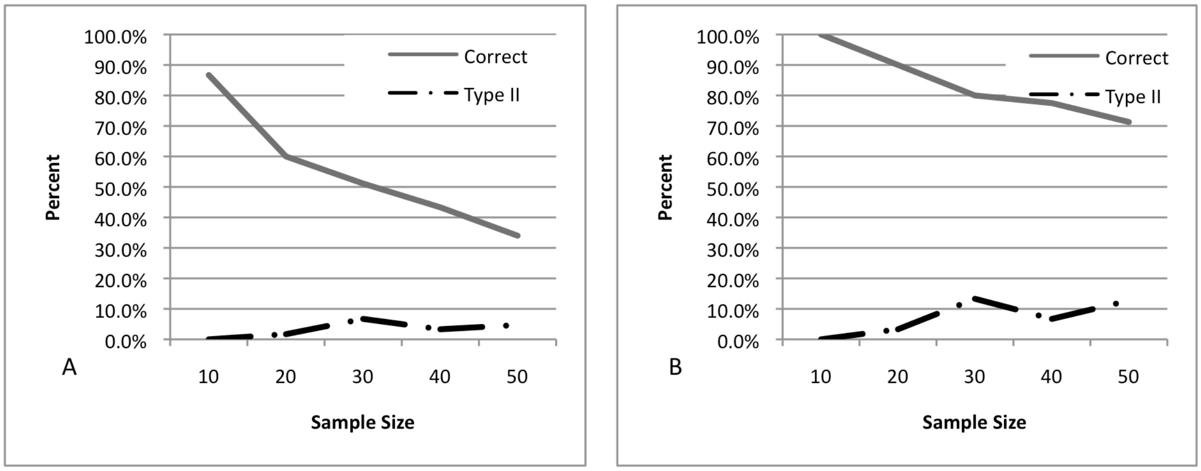

Figure 2: Graphs show the average percentage of refit success rates and Type II errors for A) P_{min} and B) P_{max}

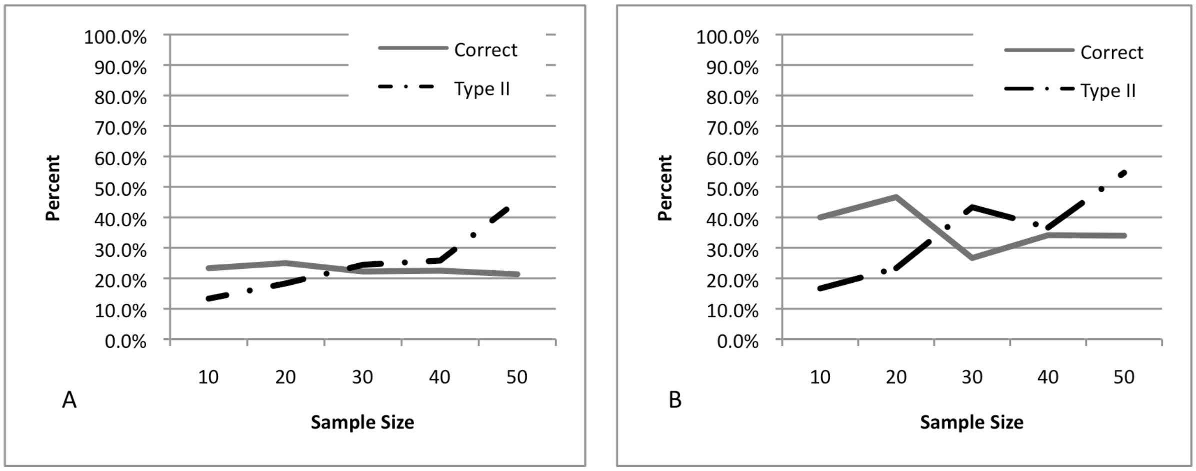

Figure 3: Graphs show the average percentage of refit success and Type II errors for Lyman's method: A) Conservative Threshold Value of 0.36 and B) Liberal Threshold Value of 0.52.